1. Données

Nous traitons le fichier « **seeds.xlsx** ». L'objectif est de détecter la variété de graines de blé (WHEAT) à partir de leurs caractéristiques (AREA … GROOVE).

2. LDA et sélection de variables sous TANAGRA

[TUTO 1] http://tutoriels-data-mining.blogspot.fr/2008/04/analyse-discriminante-linaire.html

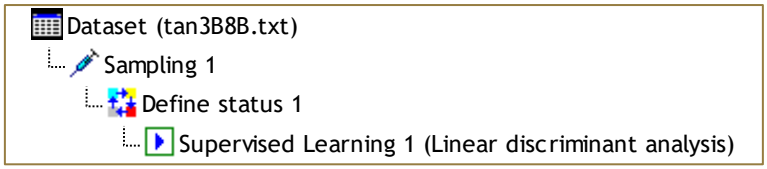
[TUTO 2] http://tutoriels-data-mining.blogspot.fr/2012/07/analyse-discriminante-lineaire.html

[TUTO 3] http://tutoriels-data-mining.blogspot.fr/2008/03/stepdisc-analyse-discriminante.html

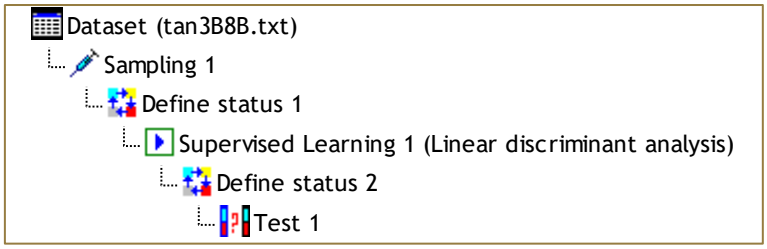
Deux notions clés supplémentaires seront abordés dans cet exercice : scinder les données en échantillons d'apprentissage et de test dans le logiciel ; procéder à une sélection de variables.

A rendre : un rapport HTML généré par Tanagra (menu DIAGRAM / CREATE REPORT) (vous les réunissez dans un seul fichier archive ZIP).

1. Importez le fichier « **seeds.xlsx** » dans Tanagra.
2. Partitionnez les données en échantillon d'apprentissage (sélectionné) et test (non sélectionné) à l'aide du composant SAMPLING (onglet INSTANCE SELECTION).
3. Créez le modèle prédictif à l'aide de LINEAR DISCRIMINANT ANALYSIS (onglet SPV LEARNING), après avoir précisé WHEAT comme variable cible (TARGET), les autres comme prédictives (INPUT) à l'aide du composant DEFINE STATUS [TUTO 1]. Votre diagramme devrait se présenter comme suit

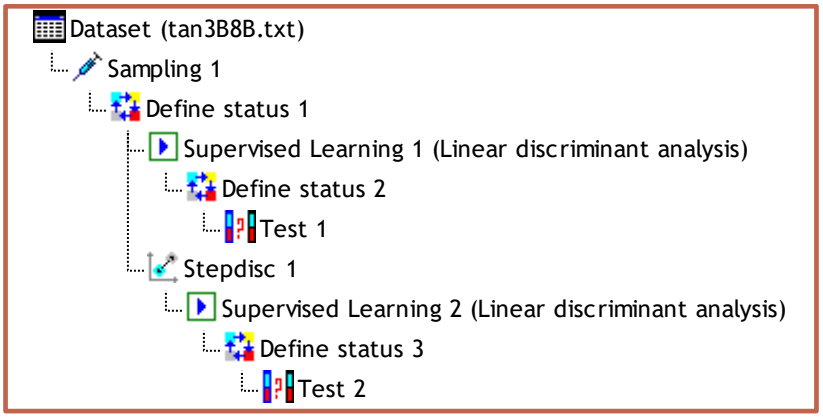


4. Nous souhaitons évaluer les performances du modèle sur l'échantillon test (les individus non sélectionnés par SAMPLING). Nous insérons un second DEFINE STATUS pour confronter les valeurs observées de WHEAT (Target) avec celles prédites par le modèle PRED_SPVINSTANCE_1 (Input). Nous ajoutons ensuite le composant TEST (onglet SPV LEARNING ASSESSMENT) en veillant à travailler sur les individus non sélectionnés (Unselected) dans le paramétrage. Vous devriez avoir la trame suivante :



5. Prenez note de la matrice de confusion et du taux d'erreur en test. La variable cible présente 3 modalités, comment le taux d'erreur a-t-il été calculé ?
6. On souhaite procéder à une sélection de variables à l'aide de STEPDISC (onglet FEATURE SELECTION) [TUTO 3, page 6 ; le composant BOOTSTRAP ne nous concerne pas dans cet exercice]. Réalisez une sélection **FORWARD au risque 1%**. Enumérez les variables sélectionnées.
7. Enchaînez avec une analyse discriminante sur les variables sélectionnées puis évaluez de nouveau le modèle sur l'échantillon test. Quel est le taux d'erreur cette fois-ci ? Votre modèle est-il meilleur ou pire qu'à la question n°5 ? Quelle conclusion peut-on en tirer ?

Votre diagramme devrait se présenter comme suit à l'issue de l'ensemble des traitements :



3. LDA et sélection de variables sous R

Réalisez exactement la même analyse sous R. Voir [TUTO 4] http://tutoriels-data-mining.blogspot.fr/2012/07/analyse-discriminante-lineaire.html, à partir de la page 19.

A rendre : un fichier **.docx** issu d'un projet R MARKDOWN.

Remarques :

- La procédure sample() devrait vous permettre de scinder en deux data frame « apprentissage » et « test » les données importées (ex. http://tutoriels-data-mining.blogspot.com/2018/04/deep-learning-tensorflow-et-keras-sous-r.html ; page 3)
- Utilisez la commande greedy.wilks() du package « **klaR** » pour réaliser la sélection de variables.
- Les résultats peuvent être différents de ceux de Tanagra parce que les échantillons d'apprentissage et de test ne sont pas forcément identiques dans les deux outils.